KEKER VAN NEST & PETERS LATHAM & WATKINS LLP MORRISON FOERSTER

June 13, 2025

Hon. Ona T. Wang
Daniel Patrick Moynihan
United States Courthouse
500 Pearl St.
New York, NY 10007-1213

cc: All counsel of Record Line (via ECF)

Re: <u>**OpenAI's response to News Plaintiffs' motion to compel production of "datasets"**</u>
In re OpenAI Copyright Infringement Litigation, No. 1:25-md-03143-SHS-OTW
This document relates to: *NYT v. Microsoft, et al.*, No. 1:23-cv-11195-SHS-OTW

Dear Judge Wang:

Although this case is focused on ***training*** data,[1] OpenAI said it would investigate plaintiffs' requests for production of peripheral data, and has not refused to produce anything. But plaintiffs rushed to file this motion, and their haste shows. Plaintiffs demand production of data that has already been produced. They guess—incorrectly—about what they might find at various paths, leading them to misportray the relevance. Even worse, plaintiffs mischaracterize RFP scope. And despite complaints about the meet-and-confer process, 7 of 19 items (almost 40%) were requested ***four business days*** before plaintiffs filed and ***one day*** after OpenAI confirmed it would investigate. Indeed, their request for "directory information" (ECF 152 ("Mot.") at 3), was raised for the ***first time*** in the motion itself.[2]

1. Background.

At enormous expense, OpenAI has produced or made available for inspection many tens of thousands of gigabytes of text-training data, conducted searches that plaintiffs selected, and is currently conducting another round of broad searches. This motion, however, does not relate to training datasets, but instead seeks a mixture of secondary items.

Rather than serve RFPs, agree on scope, and let OpenAI conduct a reasonable search, plaintiffs comb OpenAI's technical productions for stray references that they "***believe*** . . . relate to OpenAI's efforts to rank News Plaintiffs' domains based on the quality of News Plaintiffs' content," "***believe*** contain[] a regurgitation analysis of <u>actual</u> ChatGPT user conversations," or "***believe*** include[] their content." Mot. at 2, 3 n.5 (emphases added). Plaintiffs' conjecture is mistaken.

[1] *See* ECF 72 ¶¶ 64, 77–96, 126, 128, 152, 166–70, 190–94; *Daily News v. Microsoft*, No. 1:24-cv-03285-SHS-OTW, ECF 1 ¶¶ 80–95, 192–95.
[2] OpenAI uses Plaintiffs' numbers in Exhibit A (ECF 152-2) to identify their requests. Plaintiffs indicate item 13 was raised in March (Ex. G), but it was actually June (Ex. I).

Still, OpenAI offered to investigate these requests, which cover years of historical work involving many distinct teams. Merely finding a current employee who remembers each project is burdensome, and plaintiffs inexplicably refused to share Bates numbers or file names to help identify the relevant teams. Ex. I at 1. Because engineers have done projects for OpenAI's legal team, a detailed review is needed to avoid producing data that reflect attorney-client communications, mental impressions, or material prepared in anticipation of litigation. Meanwhile, plaintiffs' requests kept changing. OpenAI produced 5 of 7 items requested in the February letter (Ex. F at 5). Plaintiffs failed to review that production, and days ago demanded many of the same items while adding many new requests. Ex. I at 5–6. OpenAI confirmed it was working with technical staff to review for responsiveness, privacy, burden, and privilege. But plaintiffs rushed forward with this motion.

2. Plaintiffs' motion is based on speculation and mischaracterization.

Had plaintiffs not rushed to Court, they would soon have learned that ***many*** of the requests have been produced, are irrelevant, are far too voluminous to produce, or are not what plaintiffs believe them to be. For example:

- Item 2 was produced months ago.
- Based on OpenAI's investigation to date, items 9–12 and 16 are not what plaintiffs "believe" and contain an estimated [redacted] of data (far more data than plaintiffs have searched in the last year). There is no practical way they could analyze so much data, so OpenAI asked to discuss what plaintiffs hoped to learn in the hopes of finding a "more convenient, less burdensome, or less expensive" source. Fed. R. Civ. P. 26(b)(2)(C)(i).
- In the few days it has had to investigate plaintiffs' 7 new requests, OpenAI has preliminarily determined:
  - Item 13 contains over [redacted] (almost entirely irrelevant image data).
  - Item 17 is ***not*** a dataset of regurgitations nor does it refer to The Times as plaintiffs "believe."
  - Item 19 points to infrastructure that hasn't existed in years.

In plaintiffs' haste to manufacture a dispute, they have attempted to shoehorn their requests into existing RFPs, but they have omitted critical limitations on scope.

- RFPs 1 and 8 were limited to ***training data*** for specific models, and since plaintiffs are not requesting training data, those RFPs are irrelevant. Ex. 1 at 11; Ex. 2 at 14.
- RFPs 25 and 33, which sought analysis and decisionmaking documents (not datasets), are inapplicable. OpenAI objected to the scope of these RFPs (Ex. 3 at 15–16, 22–23) and its July 2024 offer to meet and confer is still pending.
- Plaintiffs represent that RFPs 74, 77, and 80–81 seek "[a]ll documents relating to the NYT Datasets," but plaintiffs long-ago agreed to narrow the scope to documents "(a) ***discussing*** how and why OpenAI curates and uses different types of data to train OpenAI's relevant models, . . . and (b) ***describing*** whether and how the relevant models were trained using articles published by The Times (if at all)." Ex. 4 at 7. The data in this motion falls into ***neither*** category.
- RFP 120 requests documents concerning "output [of] copyrighted content (or suspected copyrighted content) to the user." Ex. 5 at 30. Plaintiffs "believe" that item 17 is

responsive, but [redacted] Mot. at 3.

- RFP 122 originally sought "the contents of any dataset," but plaintiffs later agreed to a narrower scope limited to "non-privileged documents . . . ***discussing or describing*** (i) techniques, technologies, or procedures . . . to prevent the relevant models from duplicating text training data; (ii) the reasons for implementing those . . . and (iii) the prevalence of outputs that are verbatim copies of news content in the text [training] data." Ex. 5 at 32–33. OpenAI stated it would search for documents "discussing or describing the [redacted] . . . project[]," but not ***every*** associated dataset. Ex. 6 at 6.

3. Plaintiffs have failed to satisfy Rule 26.

Plaintiffs have failed to meet their burden of demonstrating relevance and proportionality. *See NYT v. Microsoft*, 757 F.Supp.3d 594, 597 (S.D.N.Y. 2024) (Wang, M.J.). ***First***, plaintiffs have failed to demonstrate relevance for most of these items because they rely on conjecture and "belief." Some of their requests are more relevant, such as their request for a copy of the NYT Annotated Corpus, which Plaintiffs raised for the first time on June 4. Despite being outside the agreed scope of plaintiffs' RFPs, OpenAI offered to search for it and will produce if located. But the infrastructure where it was originally stored hasn't been used in years, and despite OpenAI's RFPs requesting plaintiffs' copy, they refuse to produce it to aid the search. Ex. I at 1, 2; Ex. 7 at 16 ; Ex. 8 at 14. Other items are also outside the scope of plaintiffs' RFPs, but OpenAI said it would investigate to determine if the data was relevant, burdensome, privileged, and/or private.

Second, plaintiffs insist on production despite failing to consider or address Rule 26 factors including benefit, undue burden, and whether information can be obtained from another source that is more convenient, less burdensome, or less expensive. Based on OpenAI's initial investigation, it is clear that Plaintiffs' requests would impose an extreme burden on OpenAI. For example, item 13 is [redacted] of data, much of it image data. Items 9–12 contain over [redacted] of data, including [redacted]. Item 17 contains [redacted]. And as OpenAI explained, some items need to be carefully reviewed for attorney-client communications, attorney impressions, or analyses prepared for litigation.

Third, over 40% of plaintiffs' requests (items 13–19) were requested ***days*** before this motion was filed. Plaintiffs have not even tried to engage in a meet-and-confer process (indeed, the parties ***never*** discussed most of those items), and on that basis alone this motion should be denied. *See* M.J. Wang Individual Practices in Civil Cases § II.b (parties shall seek relief only if "this meet-and-confer process does not resolve the dispute"). Even worse, plaintiffs used their motion to request—for the first time—"directory information." If they'd bothered to ask, they would have learned that the relevant accounts include [redacted] from an array of confidential projects. (Even with a fiber-internet connection, it would take over 80 years to download!). The federal rules do not support such extraordinary fishing expeditions.

Plaintiffs' only argument on Rule 26 is that "sheer volume alone is an insufficient reason to deny discovery." *See* Mot. at 3 (quoting *In re Adelphia Commc'ns*, 338 B.R. 546, 553 (Bankr.

S.D.N.Y. 2005)). But plaintiffs misunderstand *Adelphia*, where the party ***seeking*** discovery objected to the volume of documents made available for inspection. 338 B.R. at 549. *Adelphia* says nothing about the proportionality factors of Rule 26, such as importance, burden, expense, likely benefit, or availability of more convenient, less burdensome, or less expensive sources. It also says nothing about whether requests must be granted, even if outside the scope of RFPs, and supported only by conjecture.

Respectfully,

KEKER, VAN NEST &
PETERS LLP[3]

/s/ Thomas E. Gorman
Thomas E. Gorman

LATHAM & WATKINS
LLP

/s/ Herman H. Yue
Herman H. Yue

MORRISON & FOERSTER
LLP

/s/ Rose S. Lee
Rose S. Lee

[3] All parties whose electronic signatures are included herein have consented to the filing of this document.